ISSN: 2664-9640

Predicting the Choice of Bank (Public or Private) based on Independent Variables

Oumarou Djonret Tawa, Shepherd Chikomana *

School of Science, Zhejiang University of Science and Technology, Hangzhou, 310023, China

Abstract

This study aims to predict the odds of selection between Public and Private bank on the perception of respondents in relation to the independent factors like Value Added Services, Perceived Risk, Reputation and Perceived Costs. Binary logistic regression model was built to predict the likelihood that each interviewed respondents will choose Public and Private bank, based on independent variables. The dependent variable in this study was 'Preferred Choice of Bank [Choice]' (coded 1= Private and 0=Public), where the 'Public' group serves as the reference/baseline category and the 'Private' being the target category. All the Predictors are assumed continuous in the model and are written as Value Added Services(VAS), Perceived Risk(PR), Reputation and Perceived Costs(PC).

Keywords

Nagelkerke, dichotomous, model summary, block 0, pseudo-R-square, The Hosmer & Lemeshow.

1. Introduction

Binary logistic regression analysis is a statistical approach employed to establish the connection between a binary dependent variable (also referred to as the outcome or response variable) and one or more independent variables (also called predictors or covariates). This regression technique is utilized to predict the probability of an occurrence in numerous fields, such as healthcare, social science, and business. According to [1], binary logistic regression can be used to estimate the likelihood of a binary outcome variable, such as the likelihood of a patient having a certain disease, based on one or several independent variables, such as age, gender, and health status.

The model is called 'logistic' because it uses the logarithm of the odds of the event occurring, instead of the probability itself. The logistic regression model assumes a linear relationship between the independent variables and the logarithm of the odds of the outcome variable [2]. Model estimates coefficients for independent variables in order to find out the magnitude and direction of the effect of these variables on the outcome variable.

The coefficients of independent variables can now be used to predict the likelihood of an outcome variable, given certain values of the independent variables. One of the advantages of binary logistic regression is having ability to handle both categorical and continuous independent variables [3]. Fundamentally, the researcher is addressing the question, "What is the probability that a given case falls into one of two categories on the dependent variable, given the predictors in the model?". As one might be inclined to ask why we don't use standard ordinary least squares regression (OLS) instead of BLR, OLS regression assumes (a) a linear relationship between the independent variables and the dependent variable, (b) the residuals are normally distributed, and (c) the residuals exhibit constant variance (Pituch & Stevens, 2016)[4]. All three assumptions are violated if the outcome variable in an OLS model is binary. And pivoting off (a), the relationship between one or more predictors and the probability of a target outcome is inherently non-linear as probabilities are bounded at 0 and 1.

ISSN: 2664-9640

When modeling a binary outcome using OLS regression, the estimation of model parameters ignores this boundedness, which has the notable effect of producing predicted probabilities that fall outside the 0-1 range. BLR estimates regression parameters by considering the fact that probabilities are bounded and 0 and 1. It also does not assume that residuals are normally distributed and exhibit constant variance.

2. Literature Review

2.1. Model Estimation

BLR, in contrast to OLS regression, employs maximum likelihood (ML) to estimate the parameters of the model. Maximum likelihood estimation is an iterative procedure that aims to determine the population (parameter) values that are most likely responsible for generating the observed (sample) data. Typically, this estimation method assumes the availability of large samples, and smaller sample sizes can pose challenges regarding model convergence and parameter estimation, apart from power-related issues.

[Note: In cases of smaller samples, alternative techniques such as Exact logistic regression or Firth procedure using Penalized Maximum likelihood can be utilized. Regrettably, these options are not commonly found in statistical software.] Using our sample, the model will forecast outcomes based on the independent variables of avoidance of disclosure and symptom severity. It will also present a table that illustrates the odds ratio associated with each predictor, enabling us to draw conclusions.

2.2. Evaluation of Model Fit

There are several methods available to assess the model fit in binary logistic regression, including goodness-of-fit tests, pseudo R-squared values, and receiver operating characteristic (ROC) curves. Goodness-of-fit tests evaluate the overall fit of the model by comparing observed frequencies with the expected frequencies derived from the model. These measures provide valuable insights into the adequacy of the binary logistic regression model. With the Hosmer-Lemeshow test [5] being the most commonly used goodness-of-fit test , which divides the data into groups based on predicted probabilities and compares the observed and expected frequencies within each group. A significant result indicates poor model fit.

Pseudo R-squared values: these are measures of how well the model explains the variability in the data. The most commonly used pseudo-R-squared values are Nagelkerke's R-squared [6], Cox and Snell R-squared [7], and McFadden's R-squared [8]. These values range from 0 to 1, with higher values indicating better model fit. ROC curves: these plots show the trade-off between sensitivity (true positive rate) and specificity (true negative rate) for different classification thresholds.

The area under the curve (AUC) [9] is a commonly used measure of model fit, with values ranging from 0.5 (random guessing) to 1 (perfect classification). With that being said, It is important to note that none of these measures alone can fully capture the performance of a logistic regression model, and a combination of these measures should be used to assess the overall fit of the model.

3. Methodology

3.1. Data

This study was based on 341 interviewees who gave an honest assessment on the likelihood of choosing a public bank or a private bank. The study aims to determine this choice based on several independent variables. Value Added Services(VAS), this refers to additional services that the bank may offer, such as financial planning, investment advice, or other perks.

ISSN: 2664-9640

Reputation, this refers to the overall reputation of the bank, based on factors such as customer satisfaction, quality of service, and public perception. Perceived Costs(PC), this refers to the costs associated with using the bank, such as fees for services, account maintenance, or other expenses. Perceived Risk(PR), this refers to the level of perceived risk associated with using the bank, such as the safety and security of deposits, potential for fraud, or other risks. The dataset is available for download at data.sav - Google Drive

3.2. The Model

When dealing with a binary outcome, our goal is to estimate the likelihood (probability) of belonging to a specific outcome (Prob(Y=1; terminate early), based on the information we have from a set of predictors. Unfortunately, we cannot directly model this relationship using standard OLS regression because the relationship between the predictors and the probability of Y=1 follows a non-linear pattern (represented by an S-shaped logistic curve). In logistic regression, this issue is addressed by "linearizing" the relationship through the use of a logit link function. Consequently, the dependent variable we are directly predicting becomes logits, which are a mathematical transformation of probabilities. The logistic regression prediction equation can be expressed as:

$$\text{logit}(p) = \ln\left(\frac{P}{1-P}\right)$$

$$= \boldsymbol{B_0} + \boldsymbol{B_1} Value\ Added\ Services + \boldsymbol{B_2} Perceived\ Risk + \boldsymbol{B_3} Reputation + \boldsymbol{B_4} Perceived\ Costs$$

$$\pi(x_i) = \frac{e^{\boldsymbol{B_0}+\boldsymbol{B_1} Value\ Added\ Services+\boldsymbol{B_2} Perceived\ Risk+\boldsymbol{B_3} Reputation+\boldsymbol{B_4} Perceived\ Costs}}{1+e^{\boldsymbol{B_0}+\boldsymbol{B_1} Value\ Added\ Services+\boldsymbol{B_2} Perceived\ Risk+\boldsymbol{B_3} Reputation+\boldsymbol{B_4} Perceived\ Costs}}$$

Where logit(p) is our dependent variable terminate, B_0 is a constant term and It's calculated value from variables in the equation table is *1.719*, B_1 is the coefficient of the predictor *Value Added Services* and its calculated value from variables in the equation table is *.366*, B_2 is the coefficient of the predictor *Perceived Risk* and its calculated value from variables in the equation table is *-.478*, B_3 is the coefficient of the predictor *Reputation* and its calculated value from variables in the equation table is *-.206* and B_4 is the coefficient of the predictor *Perceived Costs* and its calculated value from variables in the equation table is *-.242*.

3.3. Assumption Checking

Before running a binary logistic regression model, there are some assumptions that are supposed to be met on the data and if all the assumptions are met, then we can analyze our data using binary logistic regression model. These includes:

The dependent*/response variable is binary or dichotomous[1].*

Dependent Variable Encoding

Original Value	Internal Value
Public	0
Private	1

Figure 1: Dependent Variable Encoding

ISSN: 2664-9640

Logistic regression assumes that the response variable only takes on two possible outcomes. From Figure 1, we can clearly see that we have two outcomes coded 0 and 1 for 'Public' and 'Private' respectfully as our only two outcomes, therefore our model meets this assumption.

The Observations *are Independent*

According to this assumption, the observations in the dataset must be unrelated and independent of each other. This means that they should not be correlated or arise from multiple measurements of the same entity. In our specific data, each client's observations are completely independent of one another.

Little or no multicollinearity between the predictor/explanatory variables[10].

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	Value Added Services	,777	1,287
	Perceived Risk	,779	1,284
	Reputation	,691	1,448
	Perceived Costs	,691	1,447

a. Dependent Variable: Preferred Choice of Bank

Figure 2: Collinearity Statistics

In binary logistic regression, multicollinearity refers to the presence of high correlations between predictor/explanatory variables. "Little or no multicollinearity" means that the variables included in the regression model are not highly correlated with each other. Figure 2 above, shows the Collinearity Statistics and since all our tolerance values are greater than 0.1 meaning our assumption is not violated and also, we can check the VIF values and see that all our predictor values are less than 10 confirming that there is no multicollinearity in our dataset.

Multicollinearity can cause problems in binary logistic regression, such as unstable or unreliable estimates of regression coefficients, inflated standard errors, and difficulties in interpreting the coefficients. Therefore, it is important to check for multicollinearity before fitting a binary logistic regression model.

The sample *size is sufficiently large (at least 10-20 observations per independent variable)[11].*

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	341	100,0
	Missing Cases	0	,0
	Total	341	100,0
Unselected Cases		0	,0
Total		341	100,0

a. If weight is in effect, see classification table for the total number of cases.

Figure 3: Case Processing Summary

Logistic regression assumes that the sample size of the dataset if large enough to draw valid conclusions from the fitted logistic regression model. As a rule of thumb, you should have a minimum of 10 cases with the least frequent outcome for each explanatory variable. Figure 3 above shows the total sample size of 341 clients which is a reasonably good sample size number.

There are no outliers

ISSN: 2664-9640

The logistic regression model operates under the assumption that the dataset does not contain any significant outliers or influential observations. By employing the Mahalanobis distance method to detect outliers, we observe that our data set lacks any outliers, as indicated by the minimum value being greater than 0.001. Thus, the assumption of the absence of outliers has been satisfied.

Having successfully tested and confirmed that our data fulfills all the necessary assumptions for applying the binary logistic regression model, we are now ready to commence our analysis.

4. Analysis and Results Discussion

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	341	100,0
	Missing Cases	0	,0
	Total	341	100,0
Unselected Cases		0	,0
Total		341	100,0

a. If weight is in effect, see classification table for the total number of cases.

Figure 4: Case Processing Summary

The first section of the output shows Case Processing Summary Highlighting the cases included in the analysis. In this study we have a total of 341 respondents as shown in Figure 4.

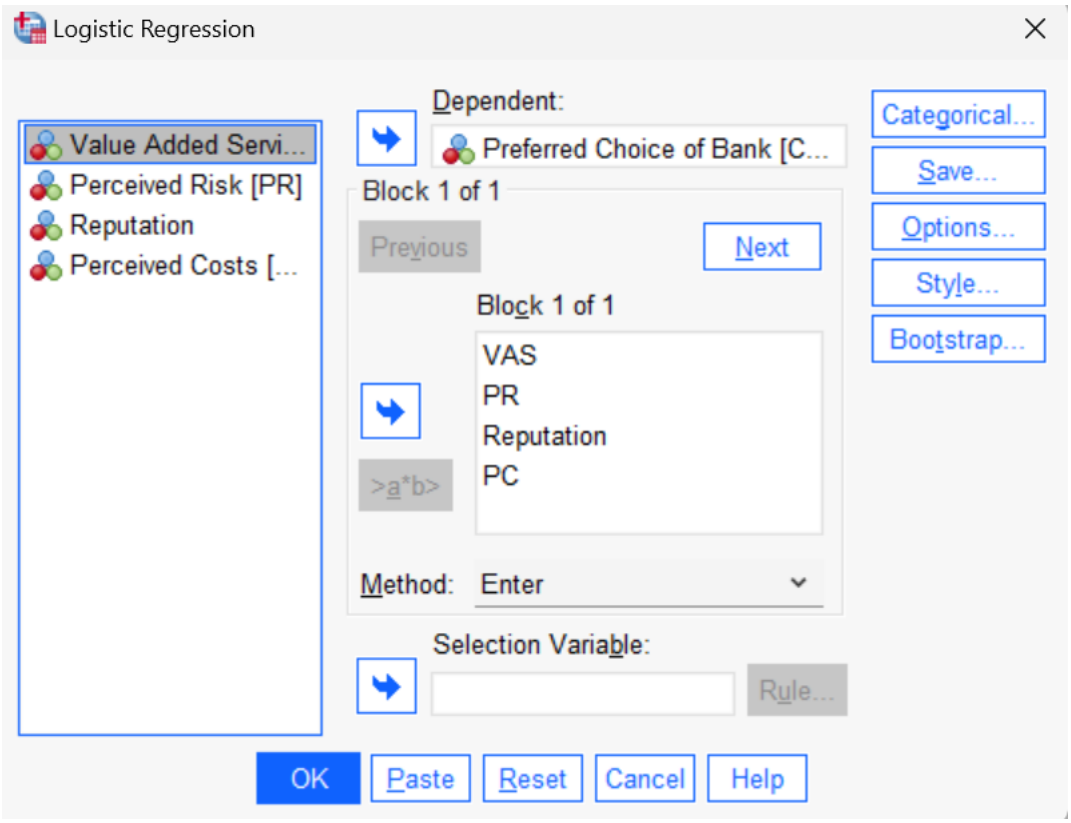

Figure 5: Logistic regression window

The diagram above (see Figure 5) shows the steps on how to execute the binary logistic regression model using SPSS.

ISSN: 2664-9640

Dependent Variable Encoding

Original Value	Internal Value
Public	0
Private	1

Figure 6: Dependent Variable Encoding

The Dependent variable encoding shows the coding for the criterion variable. In this case those who will choose Public bank are classified as 0 whereas those that choose Private bank are classified as 1(see Figure 6).

The subsequent part of the output, labeled as Block 0 in Figure 7, represents the outcome of the analysis conducted without incorporating any of our independent variables into the model. Consequently, this will serve as a reference point for comparing the model with our predictor variable included. The output is organized into blocks, with Block 0 displaying the outcomes of a null model (i.e., a model with only an intercept term).

Following Block 0, we encounter Block 1 in the output. This section is of primary importance when interpreting the results since it reflects our regression model incorporating our predictor(s).

The Omnibus Tests of Model Coefficients (see Figure 8) contains results from the likelihood ratio chi-square tests. These test whether a model including the full set of predictors is a significant improvement in fit over the null (intercept-only) model. In effect, it can be considered an omnibus test of the null hypothesis that the regression slopes for all predictors in the model are equal to zero (Pituch & Stevens, 2016)[4]. The results shown here indicate that the model fits the data significantly better than a null model, $\chi^2(4)=35.855$, $p<.001$.

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted: Preferred Choice of Bank – Public	Predicted: Preferred Choice of Bank – Private	Percentage Correct
Step 0	Preferred Choice of Bank	Public	0	90	,0
		Private	0	251	100,0
	Overall Percentage				73,6

a. Constant is included in the model.

b. The cut value is ,500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	1,026	,123	69,687	1	<,001	2,789

Figure 7: Block 0: Beginning Block

ISSN: 2664-9640

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	35,855	4	<,001
	Block	35,855	4	<,001
	Model	35,855	4	<,001

Figure 8: Omnibus Tests of Model Coefficients

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	357,745[a]	,100	,146

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Figure 9: Model Summary

The summary of the model presented in Figure 9 includes several measures: the -2 log likelihood and two 'pseudo-R-square' indices. While the pseudo-R-squares are conceptually similar to R-square, it is important to note that they are calculated differently than in OLS regression. These indices provide descriptive information and are helpful in assessing the overall adequacy of the model. The -2 Log Likelihood, also known as the model deviance, serves as an indicator of model fit. A value closer to 0 suggests a better fit, indicating less disparity between the model and the data. Conversely, higher values indicate a poorer fit, reflecting a greater difference between the model and the data. This difference in fit is manifested in the contrast between the conditional probabilities for group membership derived from the model and the actual group membership.

As noted previously, the likelihood ratio (LR) chi-square value in Figure 8 is equal to the difference in deviances (i.e., -2LL) between the model containing a complete set of predictors and reduced model containing only the intercept. We can compute the deviance for the intercept only model using the LR chi-square and the deviance of the full model:

$$Deviance(\ null\ model) = 357.745 + 35.855 = 393.6$$

The Cox & Snell and Nagelkerke R-squares are 'pseudo-R-square' values, since they are not computed the same way as R-square in the context of OLS regression. In OLS regression, R-square is interpreted as the proportion of variation accounted for in the DV as a function of the predictors. The pseudo-R-square values here are generally designed to represent proportionate change/improvement in model fit relative to the intercept-only model.

Below is the computation of Cox & Snell pseudo-R square (see Figure 10). Unlike R-square in OLS regression, the upper bound is less than 1.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	357,745[a]	,100	,146

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Figure 10: Model Summary on Cox & Snell R Square Calculation

ISSN: 2664-9640

$$CS = 1 - exp\left(\frac{deviance_{full} - deviance_{null}}{n}\right) = 1 - exp\left(\frac{357.745 - 393.6}{341}\right) = 1 - e^{-.1051}$$
$$= .100$$

*The '**exp**' in the first two expressions above are equivalents to the 'e' in third expression. This was done with the intention of enhancing the readability of the equation. Nagelkerke provided an adjustment to Cox & Snell providing an index ranging from 0 to 1. Below is the computation of the Nagelkerke pseudo-R-square.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	357,745[a]	,100	,146

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Figure 11: Model Summary on Nagelkerke pseudo-R-square

$$Nagelkerke = \frac{CS}{1 - exp\left(-\frac{deviance_{null}}{n}\right)} = \frac{.100}{1 - exp\left(-\frac{393.6}{341}\right)} = .146$$

*These versions of the Cox & Snell and Nagelkerke pseudo-R-squares are provided by Field (2018)[12]. The -2*Log likelihood (also referred to as "model deviance" [4] is most useful for comparing competing models, particularly because it is distributed as chi-square [12].

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	7,982	8	,435

Figure 12: Hosmer and Lemeshow Test

The Hosmer & Lemeshow test is another test that can be used to evaluate global fit. A non-significant test result (see Figure 12; *p=.435*) is an indicator of good model fit.

Classification Table[a]

Observed		Predicted: Preferred Choice of Bank – Public	Predicted: Preferred Choice of Bank – Private	Percentage Correct
Step 1	Preferred Choice of Bank – Public	12	78	13,3
	Preferred Choice of Bank – Private	9	242	96,4
	Overall Percentage			74,5

a. The cut value is ,500

Figure 13: Classification Table

The percentages in the first two rows provide information regarding Specificity and Sensitivity(see Figure 13) of the model in terms of predicting group membership on the dependent variable. Specificity (Also Called True Negative Rate) refers to percentage of cases observed to fall into the non-target (or reference) category (e.g., Those who will not select Private Bank) who were correctly predicted by the model to fall into that group (e.g., predicted not to select Private). The specificity for this model is $100\% \times \left(\frac{12}{12+78}\right) = 13.3\%$

ISSN: 2664-9640

Sensitivity (Also Called True Positive Rate) refers to percentage of cases observed to fall in the target group (Y=1; e.g., those who will select Private Bank) who were correctly predicted by the model to fall into that group (e.g., predicted to select Private Bank).

The sensitivity for the model is $100\% \times \left(\frac{242}{242+9}\right) = 96.4\%$.

Overall, the accuracy rate was very good, at 74.5%. The model exhibits good sensitivity since among those persons who will choose Private banks over Public Banks, 96.4% were correctly predicted to Choose Private Banks based on the model.

Within Figure 14, the 'Estimate' column presents the regression coefficients. These coefficients indicate the predicted change in the log odds of being categorized in the target group, in comparison to the reference group on the dependent variable, for every one unit increase on the respective predictor variable. This prediction is made while taking into account the effects of the other predictors in the model. Essentially, the regression slope of each predictor variable represents its impact on the log odds of the outcome variable. [Note: A common misconception is that the regression coefficient indicates the predicted change in probability of target group membership per unit increase on the predictor – i.e., p(Y=1|X's). This is WRONG! The coefficient is the predicted change in log odds per unit increase on the predictor].

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Value Added Services	,366	,101	13,063	1	<,001	1,441	1,182	1,757
	Perceived Risk	-,478	,099	23,219	1	<,001	,620	,511	,753
	Reputation	,206	,102	4,077	1	,043	1,229	1,006	1,500
	Perceived Costs	-,242	,091	7,079	1	,008	,785	,656	,938
	Constant	1,719	,623	7,609	1	,006	5,580		

a. Variable(s) entered on step 1: Value Added Services, Perceived Risk, Reputation, Perceived Costs.

Figure 14: Variables in the Equation

Nevertheless, you can generally interpret a positive regression coefficient as indicating the probability (loosely speaking) of falling into the target group increases as a result of increases on the predictor variable; and that a negative coefficient indicates that the probability (again, loosely speaking) of target membership decreases with increases on the predictor. If the regression coefficient = 0, this can be taken to indicate changes in the probability of being in the target group as scores on the predictor increase.

The Odds Ratio (OR) column contains values that are interpreted as the multiplicative change in odds for every one unit increase on a predictor. In general, an odds ratio (OR) > 1 indicates that as scores on the predictor increase, there is an increasing probability of the case falling into the target group on the dependent variable. An odds ratio (OR) < 1 can be interpreted as decreasing probability of being in the target group as scores on the predictor increase. If the OR=1, then this indicates no change in the probability of being in the target group as scores on the predictor change.

The 95% confidence interval for the Odds ratio can also be used to test the observed OR to determine if it is significantly different from the null OR of 1.0. If 1.0 falls between the lower and upper bound for a given interval, then the computed odds ratio is not significantly different from 1.0 (indicating no change as a function of the predictor).

ISSN: 2664-9640

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Value Added Services	,366	,101	13,063	1	<,001	1,441	1,182	1,757
	Perceived Risk	-,478	,099	23,219	1	<,001	,620	,511	,753
	Reputation	,206	,102	4,077	1	,043	1,229	1,006	1,500
	Perceived Costs	-,242	,091	7,079	1	,008	,785	,656	,938
	Constant	1,719	,623	7,609	1	,006	5,580		

a. Variable(s) entered on step 1: Value Added Services, Perceived Risk, Reputation, Perceived Costs.

Figure 15: Variables in the Equation Interpretation

On diagram above (see Figure 15), Value Added Services is a positive and significant (b=.366, s.e.=.101, p<.001) predictor of the probability of choosing Private bank, with the Odds Ratio indicating that for every one unit increase on this predictor, the odds of choosing Private bank change by a factor of 1.441 (meaning the odds are increasing).

Perceived Risk is a negative and significant (b=-.478, s.e.=.099, p<.001) predictor of the probability of choosing Private bank. The Odds Ratio indicates that for every one-unit increment on the predictor, the odds of choosing Private bank increase by a factor of .620 (meaning that the odds are decreasing).

Reputation is a positive and significant (b=.206, s.e.=.102, p=.043) predictor of the probability of choosing Private bank. The Odds Ratio indicates that for every one-unit increment on the predictor, the odds of choosing Private Bank increase by a factor of 1.229 (meaning that the odds are increasing).

Perceived Costs is a negative and significant (b=-0.242, s.e.=.091, p=.008) predictor of the probability of choosing Private bank. The Odds Ratio indicates that for every one-unit increment on the predictor, the odds of choosing Private Bank increase by a factor of .785 (meaning that the odds are decreasing).

5. Conclusion

Binary Logistic Regression was used to predict the choice of bank (Public or Private) based on independent variables that include Value Added Services, Perceived Risks, Reputation, and Perceived Costs. A preliminary analysis suggested that the assumption of multicollinearity was met (*tolerance=.779*). The model was statistically significant, x^2 *(4, N=341) =35.855, p<.001* indicating that the model fits the data significantly better than a null model and can model the likelihood of choosing Private bank correctly. The model explained between 10% (Cox and Snell R square) and *14.6%* Nagelkerke R square) of the variance in the dependent variable and correctly classified *74.5%* of the cases. As shown in the Table 1 below, all of our independent variables contributed to our model.

Table 1: Logistic Regression predicting the likelihood of choosing Private bank

	B	***SE***	***Wald***	***df***	***p***	***OR***	***CI for OR***	
							LL	*UL*
Value Added Services	*0.37*	*0.10*	*13.06*	*1*	*0.001*	*1.44*	*1.18*	*1.76*
Perceived Risk	*-0.48*	*0.10*	*23.22*	*1*	*0.001*	*0.62*	*0.51*	*0.75*
Reputation	*0.21*	*0.10*	*4.08*	*1*	*0.043*	*1.23*	*1.01*	*1.50*

ISSN: 2664-9640

Perceived Costs	*-0.24*	*0.09*	*7.08*	*1*	*0.008*	*0.79*	*0.66*	*0.94*
Constant	***-1.14***	***1.50***	***0.58***	***1***	***0.448***	***0.32***		

As the summary shown above (see Table 1), We can conclude that the odds of a customer choosing Private Bank offering Value Added Services are 1.44 times higher than those Public Banks which do not offer Value Added Services, with a 95% CI of 1.18 to 1.76. Also, the odds of a customer choosing Private Bank with Perceived Risk are 0.62 times lower than those Public Banks which do not have Perceived Risk , with a 95% CI of 0.51 to 0.75. Furthermore, the odds of a customer choosing Private Bank having good reputation are 1.23 times higher than those Public Banks which do not have good reputation, with a 95% CI of 1.01 to 1.50. Lastly, the odds of a customer choosing Private Bank offering perceived costs are 0.79 times lower than those Public Banks which do not offer perceived costs, with a 95% CI of 0.66 to 0.94.

References

[1] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons.

[2] Allison, P. D. (2012). Logistic regression using SAS: theory and application. SAS Institute.

[3] Menard, S. (2002). Applied logistic regression analysis (Vol. 106). Sage.

[4] Pituch, K.A., & Stevens, J.A. (2016). Applied multivariate statistics for the social sciences (6th ed). New York: Routledge.

[5] Hosmer, D.W., Hosmer, T. and Lemeshow, S. (1980) A Goodness-of-Fit Tests for the Multiple Logistic Regression Model. Communications in Statistics, 10, 1043-1069.

[6] Nagelkerke, N.J.D. (1991) A Note on a General Definition of the Coefficient of Determination. Biometrika, 78, 691-692. https://doi.org/10.1093/biomet/78.3.691.

[7] Cox, D. R., & Snell, E. J. (1989). Analysis of binary data (2nd ed.). Chapman and Hall.

[8] McFadden, D. (1974) Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka, P., Ed., Economic Theory and Mathematical Economics, Academic Press, New York, NY, 105-142.

[9] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010.

[10] Hair Jr, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). Multivariate data analysis (7th ed.). Pearson Education Limited.

[11] Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. Journal of clinical epidemiology, 49(12), 1373–1379. https://doi.org/10.1016/s0895-4356(96)00236-3.

[12] Field, A. (2018). Discovering statistics using IBM SPSS statistics (5th ed). Los Angeles: Sage.